Edcomtech: Jurnal Kajian Teknologi Pendidikan
Vol 11, No 2, Oktober 2026, pp. 339-353
https://doi.org/10.17977/um039v11i22026p339-353
pISSN: 2548-9879 (print) eISSN: 2599-2139 (online)
http://journal-fip.um.ac.id/index.php/edcomtech

Developing a Machine Learning Platform-Enhanced Deep Learning Model (MLP-DLM): Effects on Conceptual Understanding, Digital Literacy, and Higher-Order Cognitive Engagement in Islamic Boarding Schools

Muchammad Ibnu Muzakir[1*], Mahmun[2], Muhammad Romadhona Kusuma[3]
[1 2 3] *Universitas Darunnajah, Jakarta Selatan, DKI Jakarta, Indonesia*
**Email corresponding author: ibnumuzakir@darunnajah.ac.id*

Article Info

Article history:
Received 17-07-2026
Revised 17-09-2026
Accepted 01-10-2026
Published 30-10-2026

How to cite: Muzakir, M. I., Mahmun, & Kusuma, M. R. (2026). Developing a Machine Learning Platform-Enhanced Deep Learning Model (MLP-DLM): Effects on Conceptual Understanding, Digital Literacy, and Higher-Order Cognitive Engagement in Islamic Boarding Schools. *Edcomtech: Jurnal Kajian Teknologi Pendidikan, 11*(2), 339–353. https://doi.org/10.17977/um039v11i22026p339-353

© The Author(s)
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Abstrak

Pembelajaran di banyak pesantren masih berpusat pada guru, dan teknologi digital yang tersedia jarang dirancang untuk mendukung deep learning. Penelitian ini mengembangkan dan mengevaluasi Machine Learning Platform-Enhanced Deep Learning Model (MLP-DLM) yang menggabungkan pedagogi deep learning (6C), fitur machine learning platform, dan nilai-nilai inti pesantren. Pengembangan mengikuti model ADDIE. Produk divalidasi oleh tiga ahli, diujicobakan pada 30 santri, lalu dievaluasi melalui desain kuasi-eksperimen pretest-posttest dengan kelompok kontrol yang melibatkan 60 santri dari dua kelas utuh. Data dianalisis dengan uji-t, N-gain, analisis kovarians (ANCOVA), dan Cohen's d. Skor validasi ahli mencapai 4,72 dari 5,00, usability 87,5%, dan skor respons santri serta ustadz masing-masing 89,5% dan 92,8%. Dengan skor pretest yang setara, peningkatan kelompok eksperimen secara signifikan lebih tinggi daripada kelompok kontrol pada pemahaman konseptual (d = 1,83), literasi digital (d = 1,86), dan keterlibatan kognitif tingkat tinggi (d = 1,85). N-gain kelompok eksperimen tergolong sedang, sedangkan kelompok kontrol rendah. Hasil ANCOVA sejalan dengan temuan ini, tetapi untuk literasi digital dan keterlibatan kognitif tingkat tinggi hasilnya hanya diperlakukan sebagai bukti pendukung. MLP-DLM valid dan praktis, serta menjanjikan dalam konteks penelitian ini. Model ini perlu direplikasi di lebih banyak pesantren dengan intervensi lebih lama sebelum diterapkan secara luas.

Kata Kunci: *Machine Learning Platform; Deep Learning; Pemahaman Konseptual; Literasi Digital; Keterlibatan Kognitif.*

Abstract

Many pesantren (Islamic boarding schools) still rely on teacher-centered instruction, and available digital technology is rarely designed to support deep learning. This study developed and evaluated the Machine Learning Platform-Enhanced Deep Learning Model (MLP-DLM), which combines deep learning pedagogy (6C), machine learning platform features, and core pesantren values. Following the ADDIE model, the product was validated by three experts, field-tested with 30 students, and evaluated in a quasi-experimental pretest-posttest

control group design involving 60 students in two intact classes. Analyses used t-tests, N-gain, ANCOVA, and Cohen's d. The model scored 4.72 out of 5.00 in expert validation and 87.5% in usability, with response scores of 89.5% from students and 92.8% from teachers. With equivalent pretest scores, the experimental group improved significantly more than the control group in conceptual understanding (d = 1.83), digital literacy (d = 1.86), and higher-order cognitive engagement (d = 1.85). N-gain was moderate in the experimental group and low in the control group. ANCOVA supported these differences, although for digital literacy and higher-order cognitive engagement it counts only as supporting evidence. MLP-DLM is valid and practical and shows promise in this context. It should be replicated in more pesantren with longer interventions before wider adoption.

Keywords: *Machine Learning Platform; Deep Learning; Conceptual Understanding; Digital Literacy; Cognitive Engagement.*

INTRODUCTION

Pesantren have long served as a primary pillar of Islamic education in Indonesia, contributing substantially to the formation of national character and morality (Dhofier, 2011; Herningrum et al., 2020). Facing the transformation wave of the Society 5.0 era, however, the learning system in pesantren encounters substantial challenges (Badriyah et al., 2026; Mufid et al., 2026). Society 5.0 demands not only technological mastery but also deep thinking capabilities, complex problem-solving, and continuous innovation (Fukuyama, 2018). Studies nevertheless indicate that many pesantren still rely on conventional, teacher-centred methods with minimal technology utilisation to encourage deep conceptual understanding (Muid et al., 2024).

The concept of deep learning focuses on integrating new knowledge with existing cognitive structures, encouraging critical analysis, reflection, and application of knowledge in real contexts (Fullan et al., 2017; Marton & Säljö, 1976). In this article, deep learning refers to this pedagogical approach, not to the multilayer neural-network architectures used in artificial intelligence; the artificial-intelligence component of the model is the machine learning platform described below. This approach aligns with the educational goals of pesantren, which seek to form complete human beings who not only master knowledge but can also apply it wisely (Muzakir et al., 2025). Previous research has nevertheless shown a gap between deep learning potential and actual practice in pesantren (Akhirudin et al., 2025; Alfisuma et al., 2025). Learning remains dominated by memorisation and textual understanding, with limited space for exploration, collaboration, and critical thinking, which are key components of deep learning (Radino & Mubarok, 2025; Rokimin et al., 2022).

The integration of information and communication technology in pesantren education has also not been optimal. Although initiatives such as learning management systems exist, their implementation has not been structured as a driver of deep learning (Almardiah & Muis, 2025; Ningsih et al., 2025). Machine learning platforms (MLP), by contrast, offer considerable potential for transforming pesantren learning. These artificial-intelligence-based platforms provide adaptive, personalised, and interactive learning experiences through learning analytics, intelligent content recommendations, automatic feedback, and interactive simulations (Kusuma et al., 2023; Zawacki-Richter et al., 2019). Evidence indicates that constructivist and artificial-intelligence-supported technology can improve critical thinking

and conceptual understanding when it delivers personalised and actionable information (Chen et al., 2020; Hidayah & Hulyana, 2025).

A significant research gap remains: no deep learning model has been specifically designed and tested for the pesantren context using MLP. Existing models are largely generic and do not account for the culture, values, and characteristic infrastructure of pesantren. Reviews of artificial intelligence in education have similarly noted a weak connection between technological applications and pedagogical theory, together with limited attention to educators and learning contexts (Chen et al., 2020; Zawacki-Richter et al., 2019). The present study therefore develops and evaluates a contextually grounded, MLP-based deep learning model as a digital learning innovation for pesantren in the Society 5.0 era.

The research questions are: (1) What are the actual needs and challenges in implementing MLP-based deep learning in pesantren? (2) How can a prototype of MLP-based deep learning be developed that aligns with the cultural context, values, and infrastructure of pesantren? (3) How effective is the developed model in improving santri conceptual understanding, digital literacy, and higher-order cognitive engagement?

The urgency of this research is strategic on four grounds. First, pesantren must not lag behind in the digital-humanist transition. Second, the study addresses a theoretical and practical gap by producing a learning model that integrates deep learning frameworks with MLP for the pesantren context. Third, the resulting model is expected to improve learning quality and strengthen the contribution of pesantren to developing capable and principled human resources. Fourth, the initiative aligns with the national digitalisation agenda in education and the Indonesia Emas 2045 vision (Prastowo et al., 2020).

Research on technology integration in pesantren education has developed in recent years, yet fragmentation persists. Deep learning as a pedagogy has been widely elaborated (Fullan et al., 2017), but its frameworks have not been specifically adapted to the religious and cultural context of Indonesian pesantren (Hidayah & Hulyana, 2025; Muzayanah et al., 2025). Studies of pesantren digital transformation address infrastructure limitations and teacher competence (Muzakky et al., 2023), while digital literacy programmes have been shown to improve santri cognitive and practical literacy (Saepurohman et al., 2025). Work on media literacy and institutional trust has expanded pesantren reach (Bustomi et al., 2025), and digitalisation has been applied to classical learning traditions such as kitab kuning recitation (Istiqomah et al., 2025). Literature on pesantren values is likewise rich; value-based curriculum development has been examined with an emphasis on religious character formation (Azizah et al., 2023; Hamid, 2024), though without deep engagement with digital technology. Research on pandemic impacts has highlighted crisis responses but has not produced systematic digital learning models (Muzayanah et al., 2025; Yumnah, 2021).

METHOD

Research Design

This study used a research-and-development approach based on the ADDIE instructional design model (Analysis, Design, Development, Implementation, and Evaluation) (Branch, 2009). ADDIE was selected because it provides a systematic, flexible, and iterative framework for developing educational products that can be tested and disseminated. Development began with a context analysis in selected pesantren in Jakarta, using purposive sampling that considered pesantren typology (salaf, modern, integrative) and infrastructure readiness. Needs data were obtained through classroom observation, interviews with kiai, ustadz, and

santri, and review of curriculum and infrastructure documents, with source triangulation across the three evidence types (Creswell & Clark, 2017).

The design stage produced the conceptual framework of the MLP-DLM, integrating learning objectives oriented toward higher-order thinking and digital literacy, contextualised teaching materials aligned with pesantren values, learning activities that use MLP features (adaptive learning, learning analytics, intelligent content recommendations) as enablers of deep collaboration, and authentic assessment based on projects and digital portfolios. Figure 1 presents the resulting conceptual framework, and Table 3 shows how each 6C dimension was translated into learning activities. During development, a content expert, a media and technology expert, and a learning design expert evaluated the prototype using a five-point instrument. A limited trial preceded the classroom evaluation.

Machine Learning Platform Component

MLP-DLM is the established name of the instructional model, and its digital component was delivered through Deep Santri v1.0, a user-facing prototype developed by Muchammad Ibnu Muzakir (Researcher 1) and Ahmad Farhan, S.Kom. Version 1.0 is an interactive prototype. It has not been deployed as a full machine-learning application and does not use a separately trained predictive model, so its adaptive and analytic features are rule-based. The adaptive rules are transparent and depend on the learner's recent performance: three consecutive correct responses raise the difficulty by one level, and later activities stay locked until the prerequisite tasks are completed. Content recommendations draw on learning conditions represented in the prototype, including previous errors, progress, and the learner's preferred learning resource. For example, two incorrect responses on water sources trigger a recommendation for a four-minute video on the water cycle, and the learner is shown the reason for it. The ustadz dashboard presents learning analytics as class averages, normalised gain, recent engagement, current mastery or progress, repeated errors, and support needs. It can suggest a small-group halaqah, but the ustadz makes the final instructional decision. Feedback in the executable wudhu simulation follows the same approach: the simulation calculates water use, and each range of values produces its own feedback and follow-up task.

Population and Sample

The limited trial involved 30 male santri aged 12-13 years from first-year TMI at Pondok Pesantren Darunnajah, a modern pesantren. These 30 santri later formed the experimental class, so the study involved 60 unique santri across both phases. The classroom evaluation used two intact first-year TMI classes, selected purposively, with 30 male santri in the experimental group and 30 in the control group. Three ustadz took part, and the same ustadz taught both groups. The intervention ran for eight instructional meetings on wudhu. The experimental group used MLP-DLM learning activities supported by the Deep Santri prototype. The control group studied the same wudhu material through conventional instruction and wudhu practice, without adaptive pathways, learning analytics, content recommendations, or rule-based digital feedback. Conditions were assigned by intact class, so the evaluation used a quasi-experimental pretest-posttest control-group design.

Instruments and Measurement

The instruments comprised (a) an expert-validation sheet covering 43 indicators across four domains, scored from 1 to 5; (b) santri and ustadz response questionnaires scored from 1 to 4, comprising 17 and 19 items respectively; (c) a 25-item achievement test of conceptual

understanding (alpha = .88); (d) a digital literacy test adapted for the pesantren context (alpha = .86); and (e) an 18-item cognitive engagement instrument (alpha = .85) grounded in the 6Cs, following (Fredricks et al., 2004). The engagement instrument covered all six 6C dimensions, with three indicators per dimension, measured on a four-point scale. Two observers (a lecturer in Islamic education management and a senior pesantren teacher) rated four observed meetings; self-report and observation scores were combined by simple averaging (50:50). Inter-observer agreement was Cohen's kappa = .82. Three experts validated the instruments on a five-point scale with an item-retention criterion of CVI ≥ .80; a readability trial (n = 10) led to five item revisions. Cronbach's alpha was computed at pretest (.85) and posttest (.87). Digital literacy followed (van Laar et al., 2017) and recent work on artificial-intelligence literacy (Ng et al., 2021). Alpha is reported as internal-consistency evidence, not construct validity. All outcomes were scaled 0-100.

Data Analysis

Qualitative needs-assessment data were summarised thematically and triangulated across sources. Expert-validation and user-response data were analysed using means and percentages of the maximum possible score, and content validity was expressed as a content validity index. For the quasi-experiment, Shapiro–Wilk and Levene tests examined normality and homogeneity of variance. Because conditions were assigned at the class level, observations within a class may not be independent. With only one class per condition, a multilevel or cluster-level analysis was not feasible, and class membership is fully confounded with condition; the individual-level tests below therefore rest on an independence assumption that the design cannot guarantee, and this is treated as a limitation when interpreting the results. Independent-samples t-tests compared groups at pretest, to establish baseline equivalence, and at posttest. Paired-samples t-tests examined within-group change. Normalised gain was computed for each santri individually and then averaged within each group and interpreted using (Hake, 1998) thresholds: high ($g \geq .70$), moderate ($.30 \leq g < .70$), and low ($g < .30$). Analysis of covariance with pretest score as covariate was conducted as a supporting analysis, and the homogeneity-of-regression-slopes assumption was tested explicitly. Cohen's d was reported with 95% confidence intervals to communicate the magnitude of between-group differences (Cohen, 1988). The a priori criteria for model feasibility were a mean validation score of at least 4.0 and a user-response score of at least 85% of the maximum. Statistical significance was set at $p < .05$. All descriptive and inferential values reported below were computed from participant-level data, and every derived figure was reconciled against the raw score sheets before reporting.

RESULTS

Need Analysis

The needs analysis identified three connected problems: limited deep learning implementation, suboptimal use of digital technology, and the absence of an integrated learning model that accommodates pesantren cultural values. Table 1 links each finding to the corresponding design response in the MLP-DLM.

Table 1. Needs Analysis Results and MLP-DLM Design Responses

No.	Aspect	Evidence-based finding	MLP-DLM design response
1	Deep learning implementation	Learning dominated by memorisation and textual understanding, with limited	Integration of the 6Cs (critical thinking, collaboration, communication,

		exploration, collaboration, and critical thinking	creativity, citizenship, character) into learning activities
2	Digital technology utilisation	Learning management systems and digital tools present but not structured as deep learning enablers	MLP features (adaptive learning, learning analytics, intelligent content recommendations) integrated as enablers of collaboration and deep understanding
3	Pesantren value integration	No existing model integrates technology with pesantren core values	Model designed around tafaqquh fid-din, ukhuwah islamiyah, and tawazun

Design and Development Stage

The prototype comprised an MLP-DLM framework incorporating objectives oriented toward higher-order thinking and digital literacy, contextualised materials aligned with pesantren values, learning activities built on MLP features, and authentic assessment strategies. Figure 1 shows how these components relate to the needs identified in the analysis stage and to the targeted outcomes. Table 2 summarises the ratings of the three experts across 43 indicators.

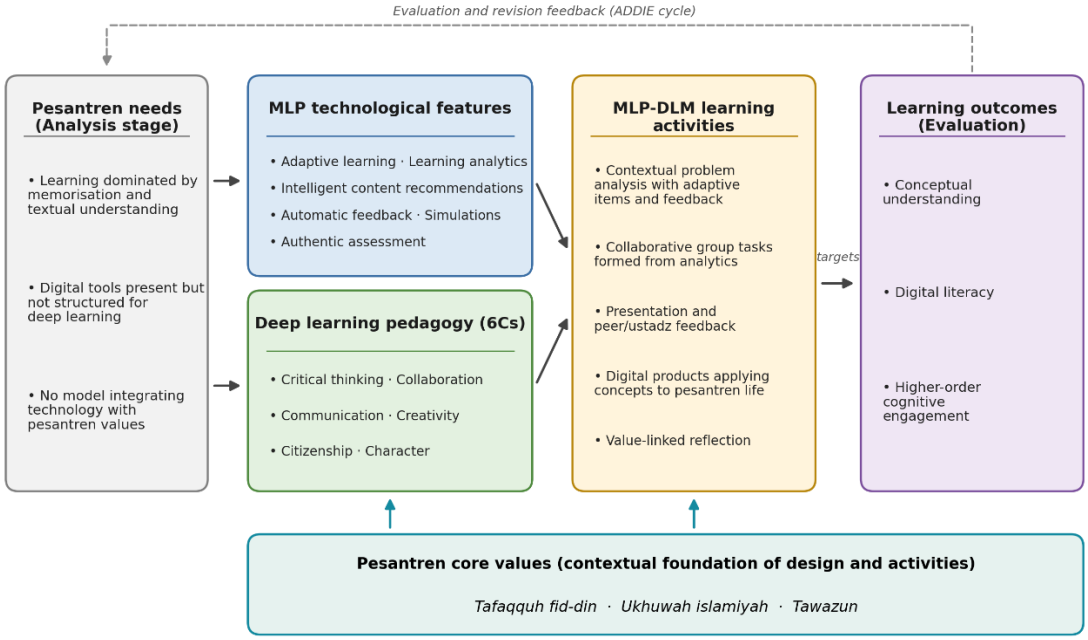


Figure 1. Conceptual Framework of the MLP-DLM

Table 2. Expert Validation Results of the MLP-DLM Model

No.	Assessment Domain	Content Expert	Media and Technology Expert	Learning Design Expert	Mean
1	Integration of deep learning principles	4.85	4.60	4.70	4.72
2	MLP feature integration and interactivity	4.55	4.80	4.65	4.67
3	Alignment with pesantren values and context	4.90	4.50	4.75	4.72
4	Assessment and evaluation design	4.70	4.65	4.80	4.72
	Expert mean	4.75	4.64	4.73	4.72

The grand mean of the domain scores was 4.72 on a five-point scale, above the predefined validity criterion of 4.00. Every indicator was rated 4 or higher, so the content validity index for that criterion was 1.00. In the limited trial with 30 santri, usability reached 87.5% of the maximum score, and the descriptive deep learning engagement indicator rose from 43.3% over the first four sessions to 76.0% over the final four. All six planned functions (adaptive learning, learning analytics, content recommendations, authentic assessment, interactive simulations, and automatic feedback) ran without a recorded system failure across 32 sessions, or 192 function-sessions in total.

Table 3 summarises how the 6Cs were operationalised in the prototype design. Each dimension was linked to a recurring learning activity, a supporting Deep Santri or MLP-DLM feature, and a pesantren value, so every C guided a specific classroom task.

Table 3. Operationalisation of the 6Cs in MLP-DLM Learning Activities

Dimension	Learning Activity	Supporting MLP Feature	Pesantren Value
Critical thinking	Santri analyse contextual problems related to the topic and justify their answers in constructed-response tasks	Adaptive learning; automatic feedback on answers	Tafaqquh fid-din
Collaboration	Small groups solve shared tasks, with roles rotated across meetings	Learning analytics used by ustadz to form mixed-ability groups	Ukhuwah islamiyah
Communication	Groups present their solutions and respond to questions from peers and the ustadz	Interactive simulations as presentation media; automatic feedback	Ukhuwah islamiyah
Creativity	Santri produce digital products (e.g., infographics or short explanations) applying concepts to pesantren life	Intelligent content recommendations of examples and resources	Tawazun
Citizenship	Projects address issues in the pesantren community, such as responsible and ethical technology use	Intelligent content recommendations of contextual materials	Ukhuwah islamiyah
Character	Santri write brief reflections linking what they learned to pesantren values	Authentic assessment through digital portfolios; progress data from learning analytics	Tawazun; tafaqquh fid-din

Meeting 6 of 8, on thaharah and wudhu, serves as an illustrative scenario of how the main platform features were intended to work. The adaptive pathway raises the difficulty level after three consecutive correct responses and keeps the next group project locked until the preceding activity is completed. After two incorrect responses about water sources, the recommendation panel suggests a four-minute video on the water cycle and explains why the video appears. The ustadz dashboard shows four santri with initial scores below 50 struggling with the same concept and suggests a small-group halaqah before the next activity opens, but the ustadz makes the final decision. The executable wudhu simulation calculates water use from duration and flow rate. It then returns rule-based feedback, ranging from efficient to excessive use, along with a follow-up task. The scenario shows the prototype's rule-based logic, and none of the functions displayed relies on a trained machine-learning model.

Baseline Equivalence of the Two Groups

Baseline equivalence was tested before outcomes were compared. The two groups did not differ significantly at pretest on any of the three outcome variables. This makes it less likely that posttest differences reflect differences in prior ability, although unmeasured class-level differences cannot be ruled out.

Table 4. Independent-Samples t-Test on Pretest Scores

Variable	Experimental M (SD)	Control M (SD)	t(58)	p	d
Conceptual understanding	56.50 (8.48)	55.17 (7.64)	0.64	.525	0.17
Digital literacy	58.63 (8.47)	56.77 (8.33)	0.86	.393	0.22
Cognitive engagement	56.13 (9.05)	55.07 (7.87)	0.49	.628	0.13

Assumption Checks

Shapiro-Wilk tests showed that all twelve distributions were normal (W = .951 to .972, p = .179 to .600). Levene's tests based on the median showed homogeneous variances at every measurement point: conceptual understanding, pretest F(1, 58) = 0.45, p = .507, and posttest F(1, 58) = 0.001, p = .970; digital literacy, pretest F(1, 58) = 0.02, p = .899, and posttest F(1, 58) = 0.68, p = .413; cognitive engagement, pretest F(1, 58) = 0.80, p = .376, and posttest F(1, 58) = 0.03, p = .862. Both assumptions were met, so no adjustment was needed.

Effects on Conceptual Understanding, Digital Literacy, and Cognitive Engagement

Table 5 reports pretest and posttest descriptive statistics, gain, percentage change, and normalised gain for the three outcome variables. Raw gains appear modest, between 5 and 17 points, but learning growth in the experimental group was 3 to 3.5 times that of the control group. Normalised gain (N-gain) helps interpret these figures. It expresses improvement as a proportion of the room each learner had to improve, meaning the distance between the pretest score and the maximum possible score, so results can be compared across groups and studies with different starting scores. After eight meetings, the experimental group reached moderate N-gain (.38 to .40), while the control group remained low (.10 to .12). These results support the conclusion that MLP-DLM produced larger learning gains than conventional instruction during the intervention.

Table 5. Changes in Conceptual Understanding, Digital Literacy, and Higher-Order Cognitive Engagement

Variable	Group	Pre M (SD)	Post M (SD)	Gain M (SD)	Change	N-gain	Category
Conceptual Understanding	Experimental	56.50 (8.48)	73.47 (7.16)	+16.97 (1.77)	+30.0%	.40	Moderate
	Control	55.17 (7.64)	60.63 (6.84)	+5.47 (1.11)	+9.9%	.12	Low
Digital literacy	Experimental	58.63 (8.47)	74.63 (6.77)	+16.00 (2.39)	+27.3%	.39	Moderate
	Control	56.77 (8.33)	61.30 (7.56)	+4.53 (1.48)	+8.0%	.11	Low
Cognitive Engagement	Experimental	56.13 (9.05)	72.60 (7.15)	+16.47 (2.27)	+29.3%	.38	Moderate
	Control	55.07 (7.87)	59.43 (7.11)	+4.37 (1.16)	+7.9%	.10	Low

Paired-samples t-tests showed significant within-group improvement in all six cases. In the experimental group, conceptual understanding increased by 16.97 points, $t(29) = 52.47$, $p < .001$; digital literacy by 16.00 points, $t(29) = 36.63$, $p < .001$; and cognitive engagement by 16.47 points, $t(29) = 39.73$, $p < .001$. In the control group, the corresponding increases were 5.47 points, $t(29) = 27.08$, $p < .001$; 4.53 points, $t(29) = 16.78$, $p < .001$; and 4.37 points, $t(29) = 20.63$, $p < .001$.

Raw gains in the experimental group were 3.1 to 3.8 times those of the control group across the three measures. In percentage terms, experimental santri improved by 30.0%, 27.3%, and 29.3%, compared with 9.9%, 8.0%, and 7.9% in the control group. The two groups had equivalent pretest scores and were taught by the same three ustadz over eight meetings. The planned difference between them was access to MLP-DLM features: adaptive pathways, learning analytics, content recommendations, and rule-based feedback. Based on Hake's (1998) criteria, normalised gain was moderate in the experimental group ($g = .38$ to $.40$) and low in the control group ($g = .10$ to $.12$). Table 6 reports the between-group posttest comparisons and their effect sizes.

Table 6. Independent-Samples t-Test on Posttest Scores

Statistic	Conceptual Understanding	Digital Literacy	Cognitive Engagement
Mean difference	12.83	13.33	13.17
95% CI of the mean difference	[9.22, 16.45]	[9.62, 17.04]	[9.48, 16.85]
Pooled standard deviation	7.00	7.18	7.13
t(58)	7.10	7.20	7.15
p	< .001	< .001	< .001
Cohen's d	1.83	1.86	1.85
95% CI of d	[1.23, 2.44]	[1.25, 2.46]	[1.24, 2.45]

The standardised between-group differences are very large ($d = 1.83$-1.86), and their confidence intervals are wide (Table 6). These estimates describe the contrast between two intact classes over eight meetings and should not be read as a generalisable estimate of the model's effect. With one class per condition, class-level influences such as timetable, peer dynamics, or class climate cannot be separated from the intervention. Effect sizes from small samples can also be unstable. The same three ustadz taught both conditions, which reduces teacher-identity differences between groups but does not remove possible class-level or treatment-diffusion effects. More precise estimates require trials with more classes per condition.

Analysis of covariance with pretest score as covariate is retained as a supporting analysis. As currently reported, conceptual understanding had adjusted means of 72.89 and 61.21, a difference of 11.69 points, $F(1, 57) = 2260.07$, $p < .001$, partial eta-squared = .975. Digital literacy had adjusted means of 73.85 and 62.08, a difference of 11.77 points, $F(1, 57) = 969.38$, $p < .001$, partial eta-squared = .944. Cognitive engagement had adjusted means of 72.16 and 59.88, a difference of 12.28 points, $F(1, 57) = 1851.42$, $p < .001$, partial eta-squared = .970. The homogeneity-of-regression-slopes assumption was reported as met for conceptual understanding, $t(56) = -1.81$, $p = .076$, but not for digital literacy, $t(56) = -2.71$, $p = .009$, or cognitive engagement, $t(56) = -3.77$, $p < .001$. Accordingly, ANCOVA for the latter two outcomes should be treated only as supporting evidence.

User response was consistently positive. Across 17 items the santri response score reached 89.5% of the maximum, distributed across ease of use (88.0%), usefulness (90.5%), and satisfaction (89.8%). Across 19 items the ustadz response score reached 92.8%, distributed across ease of implementation (90.8%), instructional effectiveness (94.5%), fit with the pesantren context (90.8%), and satisfaction and sustainability (95.0%). At feature level, adaptive learning received the strongest santri rating (94%), followed by learning analytics (92%), intelligent content recommendations (87%), and authentic assessment (83%). The lowest-rated item among ustadz concerned support for kitab kuning learning (79.2%), echoing the content expert's recommendation to add more classical-text applications. These figures should be read as response scores rather than as the proportion of respondents holding an identical opinion.

DISCUSSION

This study provides initial evidence of the MLP-DLM's promise within the studied context. The moderate gains (g=.39-.40) and large effect sizes (d=1.83-1.86) come from a single modern pesantren with limited sample size. The width of confidence intervals [1.23-2.44] reflects sampling uncertainty. Replication across multiple pesantren, class repetitions, and longer interventions is essential before claiming causal effectiveness or broader applicability.

The study produced three main findings. First, the MLP-DLM met the predetermined validity and practicality criteria after development through the five ADDIE stages. Second, the experimental group showed substantially greater improvement, associated with the MLP-DLM, in conceptual understanding, digital literacy, and higher-order cognitive engagement than the control group; covariance adjustment supported this pattern for conceptual understanding and served only as supporting evidence for the other two outcomes. Third, both santri and ustadz rated the model highly on usability, usefulness, and sustainability. Together the findings connect product development with measurable classroom outcomes.

The expert-validation mean of 4.72 and the strong user-response scores are consistent with earlier educational media and model development studies in which interactive learning products were judged valid and practical (Almardiah & Muis, 2025; Ningsih et al., 2025). The present study complements that work by examining classroom outcomes alongside validity and practicality. This matters in the study context because the needs analysis showed that digital infrastructure is increasingly available while pedagogical integration remains limited, the same pattern of technology outpacing pedagogy that reviews of artificial intelligence in education have documented more broadly (Chen et al., 2020; Zawacki-Richter et al., 2019).

Interpreting Gains Within Context

Readers familiar with educational research will note that absolute gains of 16–17 points (on a 0–100 scale) and N-gain values of .38–.40 are moderate rather than high. This is appropriate for an 8-meeting intervention and reflects realistic rather than inflated effect sizes. Normalised gain is critical here: it quantifies the proportion of available learning "headroom" (distance from pretest score to maximum possible score) that was captured. An experimental N-gain of 0.40 indicates that participants achieved 40% of the possible improvement from their pretest scores toward the maximum attainable score, representing a substantial gain within the study period. By contrast, the control group's N-gain of .12 indicates they closed only 12% of available gains, suggesting routine instruction offers minimal leverage for learning in this context. The 3-fold difference in raw gains (16.97 vs 5.47, 16.00 vs 4.53, 16.47 vs 4.37 across measures) and 3.3-fold difference in N-gains (.40 vs .12) jointly

demonstrate that the MLP-DLM substantially outperformed conventional pedagogy despite both serving equivalent populations with equivalent instructional time.

The moderate normalised gains in the experimental group (g = .40, .39, and .38) sit well above the low gains of the control group (g = .12, .11, and .10) and correspond to large standardised between-group differences (d = 1.83, 1.86, and 1.85). The magnitude of the between-group contrast rather than the absolute gain carries the argument here: an eight-meeting intervention would not be expected to produce high normalised gain in (Hake, 1998) terms, and the moderate value obtained is a realistic outcome for an intervention of this length. The result is compatible with deep learning theory, which emphasises purposeful integration of new knowledge with existing cognitive structures, critical analysis, and application in real contexts (Fullan et al., 2017; Marton & Säljö, 1976). It is also consistent with evidence that adaptive learning and intelligent content recommendations improve learning when they deliver personalised and actionable information (Chen et al., 2020; Hidayah & Hulyana, 2025). In this intervention, MLP features made learning paths visible, provided real-time feedback, and allowed personalised content delivery during the lesson. Instructionally, the 30.0% gain in conceptual understanding, against 9.9% in the control group, is consistent with the model's emphasis on explaining and applying concepts rather than recalling them. Santri in the experimental class spent a larger share of each lesson working on items matched to their current level and revising answers after feedback, whereas the control class followed a single whole-class sequence The test scores alone, however, cannot show that understanding became qualitatively deeper. The design cannot, however, isolate which component (adaptive learning, learning analytics, intelligent recommendations, or teacher training) accounted for the observed effect.

The 27.3% increase in digital literacy is plausible in light of the digital literacy literature. Programmes of this kind have been shown to improve santri cognitive and practical literacy (Saepurohman et al., 2025), and the twenty-first-century digital skills framework locates such gains in authentic, task-embedded technology use rather than in isolated technical training (van Laar et al., 2017). Working with an artificial-intelligence-driven platform may also have contributed to the emerging competencies described in the artificial-intelligence literacy literature (Ng et al., 2021). Digital literacy development nonetheless depends on infrastructure readiness, teacher competence, and alignment with pesantren cultural values (Muzakky et al., 2023). In the MLP-DLM design, santri engage with digital representations, recommendations, progress information, simulations, and digital-portfolio tasks as part of subject learning. The findings therefore support the combined instructional design used in this study rather than an unconditional claim that a machine-learning algorithm alone improves digital literacy.

The 29.3% increase in higher-order cognitive engagement is educationally important because deep learning emphasises the 6Cs (Fullan et al., 2017), and because cognitive engagement is understood as an investment in effortful, strategic learning rather than mere participation (Fredricks et al., 2004). The finding extends deep learning evidence from general education settings to the pesantren context. Deep learning activities, adaptive challenges, and visible progress may have strengthened engagement and reduced passive learning tendencies (Radino & Mubarok, 2025; Rokimin et al., 2022). Because engagement was measured through observation and self-report over eight meetings, the result should be interpreted as a short-term change rather than as evidence of durable transformation in learning habits.

One unanticipated result warrants cautious attention. The significant group × pretest interaction on digital literacy and cognitive engagement suggests that the advantage of the

MLP-DLM was larger for santri who began with lower scores, whereas on conceptual understanding the effect was uniform across the ability range. Adaptive sequencing and automatic feedback plausibly benefit learners furthest from mastery, since these are precisely the learners for whom undifferentiated whole-class instruction offers least support. This pattern points to a potential equity-related benefit for pesantren that serve heterogeneous intakes, but it should be regarded as exploratory: it was not hypothesised in advance, it comes from two classes, and it may partly reflect regression toward the mean or a ceiling on gains among higher-scoring santri. Replication with larger samples and preregistered moderation analyses is required before the model can be described as narrowing achievement gaps. The interaction simultaneously qualifies the covariance analysis, whose homogeneity-of-slopes assumption it violates, which is why the independent-samples comparison of posttest scores is treated as the primary analysis for those two outcomes.

Beyond model development, the study contributes classroom evaluation evidence. Reviews of technology in pesantren education have identified rapid growth alongside limited theoretical integration and a need for implementation research that accounts for teachers and learning contexts (Badriyah et al., 2026; Muid et al., 2024). This study responds by combining validated model development with classroom evaluation in Islamic boarding schools, and by reporting model quality, conceptual understanding, digital literacy, cognitive engagement, and user response within a single evaluation.

Several limitations bear directly on interpretation. First, the sample comprised 60 male santri in two intact first-year TMI classes from one modern pesantren, with one class per condition. Class membership is therefore fully confounded with condition, and individual-level tests rely on an independence assumption that the design cannot guarantee. The same three ustadz taught both conditions, reducing teacher-identity differences but creating a possible risk of treatment diffusion. Second, the 30 santri in the limited trial were the same 30 santri later included in the experimental class, so the development and evaluation phases were not independent and familiarity with the prototype may have influenced later performance or responses. Third, the intervention lasted eight meetings and no delayed posttest was administered, so the findings address short-term change only. Fourth, Deep Santri v1.0 was an interactive prototype rather than a fully deployed machine-learning application; only selected functions were executable in the prototype file, and the study therefore evaluates the MLP-DLM instructional package and platform logic rather than a trained predictive algorithm. Fifth, implementation fidelity was not quantified, so it cannot be determined how consistently individual features and 6C activities were delivered or which component contributed most. Sixth, cognitive engagement and practicality relied partly on self-report, and the achievement test was scored by the research team rather than independent raters. Finally, the paired-change and covariance statistics require final reconciliation against participant-level raw data before submission. Taken together, these constraints mean that the MLP-DLM should be regarded as a promising model supported by initial evidence from the participating classes. Future research should involve multiple pesantren and multiple classes per condition, independent limited-trial samples, cluster-aware analysis, delayed posttests, fidelity measures, independent scoring, and a fully implemented platform if machine-learning functionality is to be evaluated directly.

CONCLUSION

The MLP-Enhanced Deep Learning Model developed through the ADDIE model met the study's validity criterion (M = 4.72 of 5.00, content validity index = 1.00) and its practicality

criteria (usability 87.5%, santri response 89.5%, ustadz response 92.8% of the maximum). In the quasi-experimental evaluation of 60 male santri from two intact first-year TMI classes at one modern pesantren, the groups were equivalent at pretest and the experimental class subsequently achieved higher posttest scores in conceptual understanding, digital literacy, and higher-order cognitive engagement, with moderate normalised gains compared with low gains in the control class. These findings indicate greater improvement associated with the MLP-DLM instructional package in the studied classes, not established causal effectiveness or generalisability. Deep Santri v1.0 should be understood as an interactive prototype that operationalises adaptive pathways, learning analytics, transparent recommendations, and rule-based feedback; the present version does not provide evidence of a separately trained machine-learning engine. Further validation should use more pesantren and classes, an independent limited-trial sample, longer follow-up, fidelity measures, verified participant-level statistics, and a fully implemented digital platform before large-scale adoption or claims about machine-learning effectiveness are made.

REFERENCES

Akhirudin, M. T., Rahman, W., Muzakir, M. I., Ardiansyah, A., & Fauzi, A. (2025). Implementing Character Education In Islamic Boarding Schools: A Cfir Analysis of the Nisaiyah Program at An-Nur Darunnajah 8, Indonesia. *Hijri*, *14*(2), 277-288. http://dx.doi.org/10.30821/hijri.v14i2.26115

Alfisuma, M. Z., Pujiati, T., Rifa'i, B., Arjulayana, A., & Daeli, M. (2025, August). Optimizing ict integration for english learning in indonesian islamic boarding schools. In *2025 4th International Conference on Creative Communication and Innovative Technology (ICCIT)* (pp. 1-7). IEEE.

Almardiah, D. H., & Muis, A. A. (2025). The effectiveness of digital media in learning Islamic Religious Education (PAI) in the era of Society 5.0: Study of the integration of technology and religious values. *Jurnal Eduslamic*, *3*(1), 45-55. https://doi.org/10.59548/jed.v3i1.463

Azizah, N., Nurdianzah, E., Wijaya, M. M., Azami, T., & Rohman, A. (2023). Religious moderation in the industrial era 4.0: Deradicalization through the development of intellectual traditions at Fadhlul Fadhlan Islamic Boarding School Semarang. *Jurnal Pendidikan Agama Islam*, *20*(2), 233-246. https://doi.org/10.14421/jpai.v20i2.7771

Badriyah, L., & Wardi, M. (2026). Policy analysis of Islamic educational institutions in facing the challenges of society 5.0: Innovation, learning, and technology-based infrastructure. *Multidisciplinary Reviews*, *9*(1), 2026012. https://doi.org/10.31893/multirev.2026012

Branch, R. M. (2009). *Instructional design: The ADDIE approach*. Springer. https://doi.org/10.1007/978-0-387-09506-6

Bustomi, A. A., Saiban, K., Rozikin, Z., Suadi, S., & Armiah, A. (2025). Media literacy, digital exposure, governance and trust in Pesantren sustainability: the moderating role of management openness in the post-truth era. *Management & Accounting Review (MAR)*, *24*(2), 417-446. https://doi.org/10.24191/MAR.V24i02-16

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and education: Artificial intelligence*, *1*, 100002. https://doi.org/10.1016/j.caeai.2020.100002

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Dhofier, Z. (2011). *Tradisi pesantren: Studi pandangan hidup kyai dan visinya mengenai masa depan Indonesia*. LP3ES.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59-109. https://doi.org/10.3102/00346543074001059

Fukuyama, M. (2018). Society 5.0: Aiming for a new human-centered society. *Japan spotlight*, *27*(5), 47-50.

Fullan, M., Quinn, J., & McEachen, J. (2017). *Deep learning: Engage the world change the world*. Corwin Press.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American journal of Physics*, *66*(1), 64-74. https://doi.org/10.1119/1.18809

Hamid, A. (2024). Model Manajemen Kurikulum Berbasis Karakter Di Pesantren: Implikasinya Terhadap Pembentukan Kepribadian: Model Manajemen Kurikulum Berbasis Karakter di Pesantren, Implikasi terhadap Pembentukan Kepribadian Mahasiswa. *NAAFI: Jurnal Ilmiah Mahasiswa*, *1*(1), 58-64. https://doi.org/10.62387/naafijurnalilmiahmahasiswa.v1i1.80

Herningrum, I., Alfian, M., & Putra, P. H. (2020). Peran pesantren sebagai salah satu lembaga pendidikan Islam. *Islamika: Jurnal Ilmu-Ilmu Keislaman*, *20*(02), 1-11. https://doi.org/10.32939/islamika.v20i02.582

Hidayah, B., & Hulyana, I. (2025). Strategi pendekatan pembelajaran deep learning dalam meningkatkan kompetensi berpikir kritis dan kreatif di era digital. *Jurnal Ilmu Manajemen Pendidikan*, *4*(2), 113-121. https://doi.org/10.52431/manajeria.v4i02.4004

Istiqomah, N., Nadhiroh, W., Zulaichah, S., & Amnesti, S. K. W. (2025). Digitalisasi Pengajian Kitab Kuning Di Pondok Pesantren As-Sa'idiyyah 2 Bahrul Ulum Tambak Beras Jombang. *Jurnal Lektur Keagamaan*, *23*(1), 342-373. https://doi.org/10.31291/jlka.v23i1.1358

Kusuma, M. R., Gata, W., Kurniawan, S., Saputra, D. D., & Panggabean, S. (2023). Software defect prediction for quality evaluation using learning techniques ensemble stacking. *Inspiration: Jurnal Teknologi Informasi Dan Komunikasi*, *13*(2), 1-13. https://doi.org/10.35585/inspir.v13i2.58

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I—Outcome and process. *British journal of educational psychology*, *46*(1), 4-11. https://doi.org/10.1111/j.2044-8279.1976.tb02980.x

Mufid, A., Suryadi, K., Rahmat, R., & Mahpudz, A. (2026). *Defending the nation in the era of disruption: Building the character of Gen Z students through technology and character education. Multidisciplinary Science Journal, 8* (9). https://doi.org/10.31893/multiscience.2026587

Muid, A., Arifin, B., & Karim, A. (2024). Peluang dan tantangan pendidikan pesantren di era digital (Studi kasus di Pondok Pesantren Al-Islah Bungah Gresik). *MODELING: Jurnal Program Studi PGMI*, *11*(1), 512-530. https://doi.org/10.69896/modeling.v11i1.2254

Muzakir, M. I., Akhirudin, M. T., Idham, I., & Samiyono, S. (2025). Integration of Deep Learning As An Educational Innovation Strategy For Islamic Boarding Schools (Pesantren) In The Digital Era. *Jurnal As-Salam*, *9*(2), 129-141. https://doi.org/10.37249/assalam.v9i2.952

Muzakky, R. M. R., Mahmuudy, R., & Faristiana, A. R. (2023). Transformasi pesantren menghadapi era revolusi digital 4.0. *ALADALAH: Jurnal Politik, Sosial, Hukum Dan Humaniora*, *1*(3), 241-255. https://doi.org/10.59246/aladalah.v1i3.371

Muzayanah, U., Maknun, M. L., Sa'ad, F., & Taruna, M. M. (2025). Digital learning models: experience of online learning during the pandemic. *International Journal of Evaluation and Research in Education*, *14*(2), 1196-1206. http://doi.org/10.11591/ijere.v14i2.30032

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and education: Artificial intelligence*, *2*, 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ningsih, S., Suhardi, M., Sari, I. N., Wulan, D. S., Fadillah, M., & Amalia, N. (2025). Implementasi learning management system (LMS) di lingkungan pendidikan tinggi. *EDUTECH: Jurnal Inovasi Pendidikan Berbantuan Teknologi*, *5*(2), 422-432. https://doi.org/10.51878/edutech.v5i2.6370

Prastowo, A. I., Firman, A. J., Mulyanto, T., & Wiranata, R. R. S. (2020, September). The independent learning curriculum concept of Imam Zarkasyi's perspective in Pesantren for facing the era of society 5.0. In *Proceedings of the 4th International Conference on Learning Innovation and Quality Education* (pp. 1-6). https://doi.org/10.1145/3452144.3452147

Radino, R., & Mubarok, M. H. (2025). Critical pedagogy and dialogic learning in classical Islamic boarding schools: An analysis of discussion practices in Madrasah Diniyyah Al-Munawwir Krapyak. *Jurnal Pendidikan Agama Islam*, *22*(1), 255-274. https://doi.org/10.14421/jpai.v22i1.10801

Rokimin, Rofiq, M., & Rochimah, A. N. (2022). Konsep pendidik dalam perspektif Al-Qur'an (Studi kasus di Pondok Pesantren Darunnajah Jakarta). *Edukasiana: Journal of Islamic Education, 1*(1), 29–44. https://doi.org/10.61159/edukasiana.v1i1.4

Saepurohman, A., Badrudin, B., Erihadiana, M., Lestari, A. S., & Alai, A. (2025). Strategic Management of Digital Literacy Initiatives in Islamic Boarding Schools of Tasikmalaya. *Munaddhomah: Jurnal Manajemen Pendidikan Islam*, *6*(2), 238-250. https://doi.org/10.31538/munaddhomah.v6i2.1711

Van Laar, E., Van Deursen, A. J., Van Dijk, J. A., & De Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in human behavior*, *72*, 577-588. https://doi.org/10.1016/j.chb.2017.03.010

Yumnah, S. (2021). E-learning based islamic religious education of learning media: alternative solutions for online learning during covid-19. *Nazhruna: Jurnal Pendidikan Islam*, *4*(2), 249-260. https://doi.org/10.31538/nzh.v4i2.1209

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *International journal of educational technology in higher education*, *16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0